

Submission of Content to a Digital Object Repository Using a Configurable Workflow System

Andreas Hense, Johannes Müller
University of Applied Sciences Bonn-Rhein-Sieg
Grantham-Allee 20
53757 Sankt-Augustin
Germany
`andreas.hense@fh-brs.de`
`johannes.mueller@gmx.com`

February 3, 2008

The prototype of a workflow system for the submission of content to a digital object repository is here presented. It is based entirely on open-source standard components and features a service-oriented architecture. The front-end consists of Java Business Process Management (jBPM), Java Server Faces (JSF), and Java Server Pages (JSP). A Fedora Repository and a mySQL data base management system serve as a back-end. The communication between front-end and back-end uses a SOAP minimal binding stub.

We describe the design principles and the construction of the prototype and discuss the possibilities and limitations of workflow creation by administrators. The code of the prototype is open-source and can be retrieved in the project *escipub* at <http://sourceforge.net>.

1 Motivation

This work has been inspired by the eSciDoc project of the Max-Planck-Society [7]. One of the goals of the eSciDoc project is the creation of a publication management service that allows scientific organizations to establish an institutional repository.

Generally speaking, the publication process goes like this. Publications, consisting of a set of metadata and a number of content files, are submitted to a digital repository and are made publicly available following the philosophy of open access. Once publications are available they can be retrieved by a so-called persistent identifier. The organization that

2 Features of the Prototype

uses persistent identifiers guarantees that publications can always and forever be retrieved by that identifier. The responsibility for the quality of academic output, the cost of long term archival, and the fact that publications cannot be withdrawn make a well-structured quality assurance process almost mandatory.

The quality assurance process is a business process that must be adaptable to the specific needs of the organization using it. This goes beyond the mere assignment of roles to actual persons. The process must be configurable with respect to the number of steps (e. g. technical, formal, and scientific quality assurance), the degree of parallelization of different tasks, and even the quality of the tasks themselves.

The business processes of submission and quality assurance are highly structured and can be automated [6]. The part of a business process that is automated is called a workflow.

As a consequence, the software-architecture of choice for the submission of content to an institutional repository consists essentially of two parts: a workflow system and a digital object repository.

2 Features of the Prototype

The prototype is based on the *jBPM Starters Kit 3.1.1* [5] by JBoss, which has been extended by an interface to Fedora. The prototype is called "eSciPub" and can be tested under the following URL:

<http://www.bis.inf.fh-brs.de>.

The prototype features a simple role model. Users log into the system with their account and can perform tasks according to their role (author, quality assurer, process administrator,...). Every role has its proper workspace. The workspace of an author shows his pending submissions, his submissions that came back from quality assurance for rework, and a list of publication processes. The author can submit a new article by choosing one of these processes.

Every time a user is working on a task of a workflow instance, its state instance is shown in a diagram. The diagram lists the whole workflow with the current task being highlighted (cf. figure 3).

The administration of workflow instances like advancing and stopping can be performed by the process administrator using the same software as the other users. New workflows or new versions of existing ones are created graphically in the Eclipse development environment. They can also be deployed from there. Note that running workflow instances can continue with their original definition.

3 The Software Architecture

The prototype presented in this paper is built on open-source components and is itself open-source.

3.1 JBoss jBPM

jBPM [4] is a workflow management system by JBoss that is implemented in Java. jBPM has its roots in an open-source project initiated by Tom Baeyens in 2003 and which has been managed by JBoss since 2004.

The central component of jBPM is a workflow engine running processes described in XML. jBPM supports jPDL (*jBPM Process Definition Language*) and BPEL (*Business Process Execution Language*). We have chosen jPDL because BPEL was added recently as the so-called “JBoss jBPM BPEL Extension” and is still in the beta testing phase.

All relevant data for a process are stored in a single compressed file – a process archive. The process definition itself is a file in the process archive called `processdefinition.xml`. If a process definition is created by the *Graphical Process Designer* (GPD), which we will describe below, a picture of the process called `processimage.jpg` and a file containing metadata of the picture called `gpd.xml` are added to the process archive. All Java classes and libraries used by the process and all documents used as information by process owners can be added to the process archive.

Process definitions are directed graphs consisting of nodes and edges. The edges are called transitions. Nodes have a type determining their properties and behavior [3]. Every process definition has a name. XML-editors like GPD can check the validity of the jPDL-XML-structure using the XML-Namespace `urn:jbpml-3.1`.

Tasks are mapped to users or roles by the jPDL-construct *swimlane*. When the first task in a swimlane is created, the *assignment handler* of this swimlane is called; the assignment handler is defined by the attribute `assignment` of the process definition. The class referenced by the assignment handler could serve for user authentication purposes.

Process variables can be simple data types or complex Java-objects – if they are serializable. Process variables can be contained in the process definition or can be created during process execution.

Actions contain a programming logic that is executed when a certain event happens during process execution. Typical events triggering actions are arriving at a node, leaving a node, or using a transition. Actions assigned to general events and not to nodes cannot influence the control flow of the process instance. In contrast, actions assigned to nodes are responsible for the control flow of the process instance [3].

Process definitions can be created by the *Graphical Process Designer* [5], an Eclipse-plugin (cf. figure 1). Using GPD, processes can be designed graphically and be deployed on a running jBPM-server. Older versions of process definitions are not overwritten and running process instances of the older versions can be terminated using their original definition.

3.2 Fedora

The *Fedora Repository*¹ stores and manages digital objects. Fedora was selected by the eSciDoc-project as the basic technology for the institutional repository of the Max-Planck-Society [8].

¹Fedora stands for “Flexible Extensible Digital Object Repository”, cf. <http://www.fedora.info>

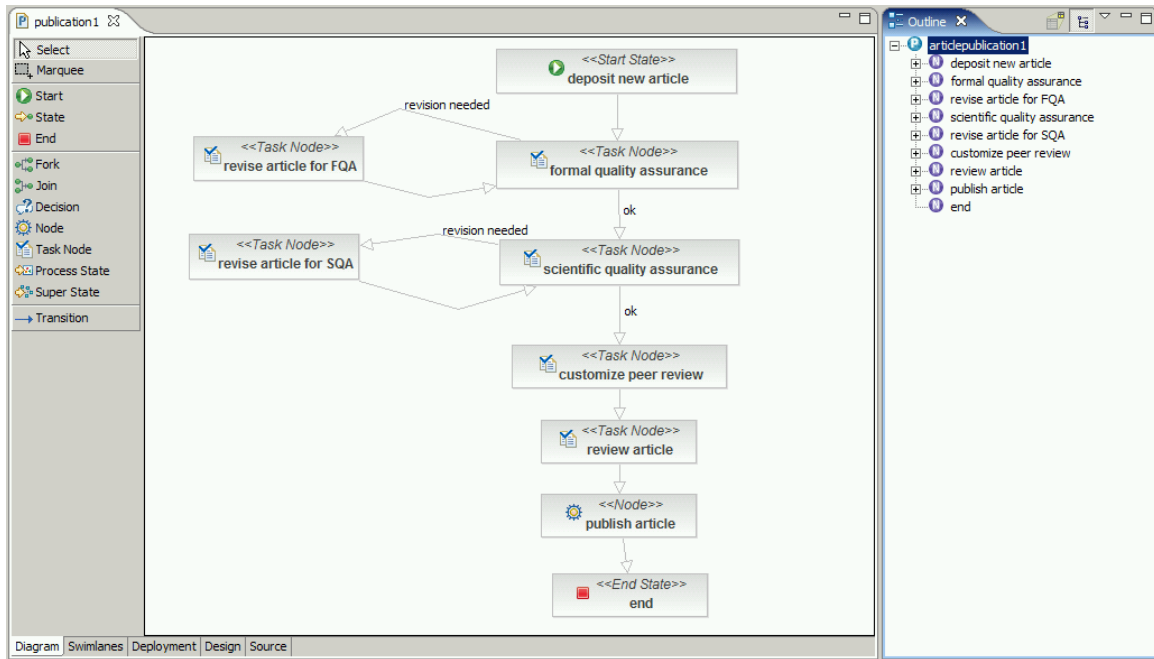


Figure 1: jBPM Graphical Process Designer (GPD).

A *Fedora-object* consists of a collection of so-called “*Content Items*” and associated meta-data. Content items contain documents, images, or any other type of file. An important feature of Fedora is that so-called “*disseminators*” can be associated with content items. This feature is essential for any scientific document server: the original \LaTeX^1 -file of this article could be contained in a content item of a Fedora-object. While the original file is only visible to the author, disseminations in HTML- or pdf-format could be presented to the public. In a slightly different context, original high definition images could be contained in content items and low resolution copies could be made publicly available by a disseminator.

Fedora is prepared for operation in a service-oriented architecture. Via the SOAP-interfaces API-A and API-M, content can be stored and retrieved in various ways. Fedora comes with an integrated Tomcat-webserver making its content accessible via HTTP. Content items can be stored locally or as references to other systems. Versioning of Fedora-objects is provided.

3.3 Further Technologies

The user interface is implemented using *Java Server Faces* (JSF) (*MyFaces* cf. <http://myfaces.apache.org>). JSF is a framework by Sun for the implementation of web applications. MyFaces is the first open-source implementation of JSF. JSF is made for processing user interactions. Its interfaces are made of elements having a state. The states of elements and events can be supervised by the JSF-instance. The tag libraries of JSF can be used in Java Server Pages (JSP). JSF runs as a servlet on the Tomcat servlet container.

¹Type setting system \LaTeX , s. <http://www.latex-project.org/>

The open-source data base management system *MySQL*¹ is used for *JBoss jBPM* and the *Fedora Repository*.

For accessing the SOAP-interface, the Apache *Axis*-library (Apache eXtensible Interaction System, cf. <http://ws.apache.org/axis/>) is used. Axis is a SOAP-engine for the construction of web services and clients.

XML-documents are constructed and accessed with the *Document Object Model* (DOM)-library of the *World Wide Web Consortium* (W3C) (cf. <http://www.w3.org/DOM/>).

The component library Apache *Tomahawk* is an extension of the MyFaces-implementation and is used for making Java Bean attributes persistent (cf. <http://myfaces.apache.org/tomahawk/index.html>).

3.4 Layers

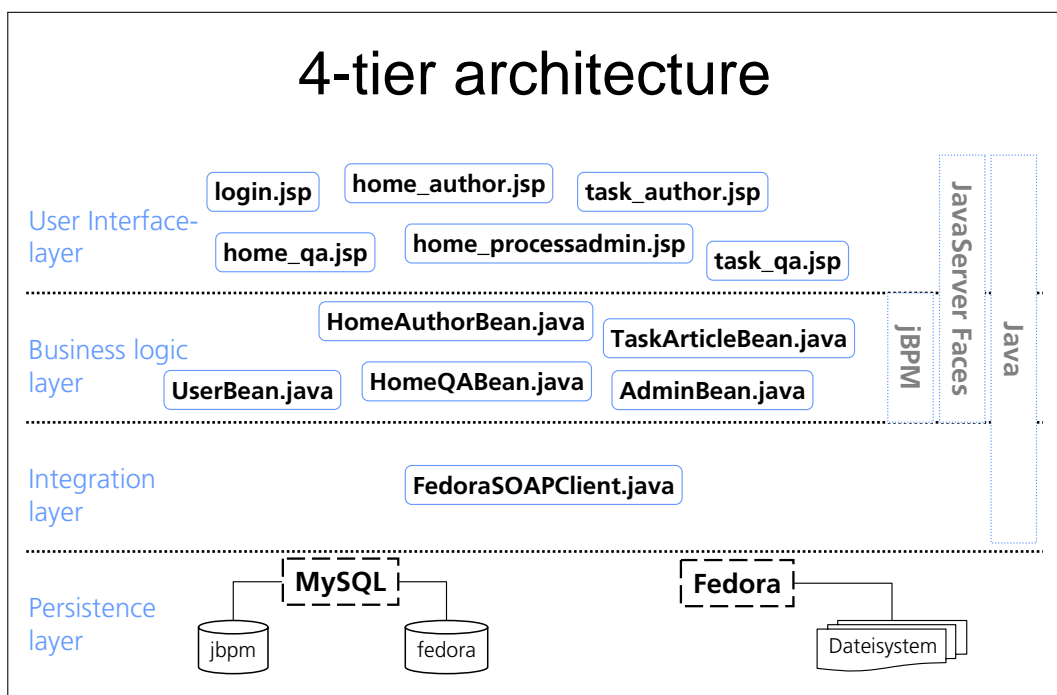


Figure 2: 4-tier architecture of the prototype.

The prototype is divided into four layers. Program modules of one layer only communicate with program modules of adjacent layers (cf. figure 2). The layers are as follows:

1. *User-Interface layer*: contains the JSP/JSF-pages for the interaction with the user.
2. *Business-logic layer*: contains the *backing beans* and Java objects implementing the “business logic”.²

¹cf. <http://www.mysql.com>

²This layer could be further divided into a layer of backing beans for the JSP and a business logic layer.

3. *Integration layer*: Objects in this layer communicate with the data stores. Their only task is taking requests from the business-logic layer and passing them to the data store – or vice versa.
4. *Persistence layer*: This layer contains no Java objects, but the MySQL DBMS and the Fedora Repository.

4 Accessing Fedora's Web Interface

4.1 Choosing the Right Way

There are several ways of accessing Fedora via its interfaces. Fedora offers a REST and a SOAP interface for calling the functions of **API-A** and **API-M**. Being the more popular protocol, we have chosen SOAP for our prototype [9]. Once the decision for SOAP is made, there are the following possibilities:

1. Using the *SOAP-Client* of the Fedora distribution [2]. This HTTP-Client is restricted to **API-A**-calls. For the functionality of our prototype, **API-M** is the more important interface. A first approach to using the program logic of the official Fedora client and extending it to **API-M**-calls was abandoned, because the Fedora client is too tightly interwoven with a number of “heavy” classes of its distribution.
2. Another approach is the direct access to Fedora using SOAP. Here, there are several sub-variants.

The variant implicitly suggested by the official Fedora documentation is the use of the Fedora client library “`client.jar`”. This library has a size of nearly one MB in version 2.1.1. It provides the data types necessary for SOAP requests to Fedora – e. g. `fedora.server.types.gen.Datastream` or `fedora.server.types.gen.RepositoryInfo`

We have chosen a more elegant - although not officially documented method: we have used a so-called *minimal binding stub* that can be generated automatically (cf. next section). Its size is 89 KB when packed as a jar-file.

Our approach follows the principle of *loose coupling* in a service-oriented architecture, because the client is independent of the “heavy” server-specific class libraries. It creates the necessary classes itself. This approach is less resource-hungry and results in a higher performance.

4.2 Generating the SOAP-Binding Stub

If programs want to use the SOAP-interface of the Fedora server, the generic data types of Fedora must be known in the runtime environment of the client program. To achieve this, there are two possibilities: include all Java classes of the Fedora implementation as source files or a jar-file, or include a *minimal binding stub*. Such a binding stub contains only those

data types that the client needs to send to or receive from the web-service-provider. As already pointed out, the last possibility is the one we prefer.

A minimal binding stub can be generated by analyzing the description of the web service. This is done by the tool “WSDL2Java” from the *Apache Axis* framework. “WSDL2Java” analyses WSDL-files (*Web Services Description Language*). WSDL-files define the signatures, i. e. the names and parameters of web services. In general, the non-primitive data types are defined in XSD-files (*XML Schema Definition*) referenced by the WSDL-file.

On the basis of WSDL- and corresponding XSD-files, WSDL2JAVA builds skeleton Java classes. The structure of the skeleton Java classes corresponds to the WSDL- and XSD-files. The skeleton Java classes contain no program logic – apart from simple getter and setter methods. Thus the client is able to call the web services and process the answers correctly.

The WSDL-files for the API-A- and API-M-SOAP-interfaces of Fedora (cf. section 3.2) can be found here:

- <http://www.fedora.info/definitions/1/0/api/Fedora-API-A.wsdl>,
- <http://www.fedora.info/definitions/1/0/api/Fedora-API-M.wsdl>.

If Fedora is installed on a standard port of the local host, the files can be found here:

- <http://localhost:8180/fedora/services/access?wsdl>,
- <http://localhost:8180/fedora/services/management?wsdl>.

Fedora-WSDL-files reference the XSD-file (*XML Schema Definition*)

<http://www.fedora.info/definitions/1/0/types/fedora-types.xsd>.

WSDL2JAVA needs the following class libraries for the creation of Java classes from Fedora-WSDL-files:

1. All program libraries of the Axis distribution, because WSDL2Java is a part of Axis.
2. With the *JavaBeans Activation Framework*¹, unknown data types can be determined, encapsulated, and their methods can be detected. The only jar-file of this distribution called `activation.jar` is needed.
3. The *JavaMail API*² is a programming interface containing a protocol independent framework for eMails and messages. The distribution contains several jar-files, of which only `mail.jar` is needed.

The XML schema definition files `fedora-types.xsd` and `fedora-auditing.xsd` must be in the path of the operating system shell in use. These files can be retrieved in the folder `xsd` or via the addresses

- <http://localhost:8180/fedora-types.xsd> or

¹cf. <http://java.sun.com/products/javabeans/jaf/>

²cf. <http://java.sun.com/products/javamail/>

- <http://localhost:8180/fedora-auditing.xsd>

of a local Fedora-server installation¹. If all the WSDL-files, XML-schema-files, and class libraries mentioned above are available, the binding stub can be generated by using the following commands:

```
java -cp C:\java\axis-1_4\lib\axis.jar;
        C:\java\axis-1_4\lib\commons-logging-1.0.4.jar;
        C:\java\axis-1_4\lib\commons-discovery-0.2.jar;
        C:\java\axis-1_4\lib\jaxrpc.jar;
        C:\java\axis-1_4\lib\saa.jar;
        C:\java\axis-1_4\lib\wsdl4j-1.5.1.jar
        C:\jaf-1.1\activation.jar;
        C:\javamail-1.4\mail.jar
        org.apache.axis.wsdl.WSDL2Java Fedora-API-A.wsdl
```

```
java -cp C:\java\axis-1_4\lib\axis.jar;
        C:\java\axis-1_4\lib\commons-logging-1.0.4.jar;
        C:\java\axis-1_4\lib\commons-discovery-0.2.jar;
        C:\java\axis-1_4\lib\jaxrpc.jar;
        C:\java\axis-1_4\lib\saa.jar;
        C:\java\axis-1_4\lib\wsdl4j-1.5.1.jar
        C:\jaf-1.1\activation.jar;
        C:\javamail-1.4\mail.jar
        org.apache.axis.wsdl.WSDL2Java Fedora-API-M.wsdl
```

At first, the generated structure of folders contains only Java source files. These have to be compiled and be put into the class path of the project as a jar-archive.

5 The Nuts and Bolts of the Web Application

One of the roles in our submission process is that of the author. He submits new content to the digital object repository. The workspace of the author (`home_author.jsp`) contains three areas: “Task-List”, “Start New Publication Process”, and an overview of all articles of this author in the repository (cf. figure 3). This section describes the mechanisms for addressing Fedora in the context of jBPM and JSF.

5.1 Display the jBPM Task List and Execute a Task

The JSF-page `home_author.jsp` with its `dataTable`-tag represents a table that is filled with the data from the `getTaskInstances`-method of the “`HomeAuthorBean`”²:

```
<h:dataTable value="#{homeAuthorBean.taskInstances}">
```

¹The URL referenced in the WSDL of the API-M-interface (<http://www.fedora.info/definitions/1/0/auditing/fedora-auditing.xsd>) was not available at the time of writing this document.

²`HomeAuthorBean.java`

It has the scope “*Request*” meaning that this bean is initialized for each request. The `JbpmContextFilter` and the constructor of the `HomeAuthorBean` ensure that the correct user- and jBPM-context-information is contained in the bean when the method is called by `home_author.jsp`.

Using the class `org.jbpm.db.TaskMgmtSession`, the function `TaskAuthorBean.getTaskInstances` can access the method `findTaskInstances`, which returns all open tasks of an actor, directly:

```
taskMgmtSession.findTaskInstances(userBean.getUserName());
```

Using the column “Name” of the task list, which contains the names of the tasks as links, the function `selectTaskInstance` of the `HomeAuthorBean` can be called:

```
<h:commandLink action="#{homeAuthorBean.selectTaskInstance}">
```

The author can submit a new article or work on an article already in the process of publication. Both of these tasks are handled by the same backing bean “`TaskAuthorBean`” (`TaskAuthorBean.java`). If the task comes from the task list, then it already exists in jBPM and does not have to be created. In this case, the `selectTaskInstance`-method of the `HomeAuthorBean` hands over the unique identifier of the selected task - the “*taskId*” - to the “`TaskAuthorBean`” (`TaskAuthorBean.java`) and displays the next page `task_author.jsp`:

```
return "task";
```

This last JSP is backed by the `TaskAuthorBean` (cf. “Submission of a new article” and “Work on an article”).

5.2 Creation of a new Fedora object

The table of process definitions on `home_author.jsp` is filled by the `getLatestProcessDefinitions`-method of the `HomeAuthorBean`:

```
<h:dataTable value="#{homeAuthorBean.latestProcessDefinitions}">
```

This method calls the method `findLatestProcessDefinitions` of the jBPM-class `org.jbpm.db.GraphSession` which returns a list of current versions of process definitions. When a new publication process is started, the method `startProcessInstance` of the `HomeAuthorBean` is called. This method creates a new Fedora object:

```
pid = fedoraSOAPClient.ingestNewFedoraObject();
```

The object creation by the “`FedoraSOAPClient`” (class `FedoraSOAPClient`) is as follows:

1. At first a new DOM-document is created as an instance of the object `org.w3c.dom.Document`:

```
Document foxmlDoc = docBuilder.newDocument();
```
2. The FOXML-root node, additional root node attributes, and the object properties *Label* and *Content Model* are added. The FOXML-DOM-document now contains the structural information necessary for insertion into the repository.
3. In order to hand over the DOM-document to Fedora by a SOAP-call, it must be serialized in a *Byte Array* first. Using the three statements

```
ByteArrayOutputStream xml = new ByteArrayOutputStream();
XMLSerializer ser = new XMLSerializer(xml, null);
ser.serialize(foxmlDoc);
```

the DOM-document `foxmlDoc` is transformed to a `java.io.ByteArrayOutputStream`.

4. The SOAP-call is prepared by creating an instance of the *Axis*-object `org.apache.axis.client.Service`:
`Service service = new Service();`
5. Using the `service`-object, the SOAP-client can create an instance of `org.apache.axis.client.Call`. The latter calls a SOAP-remote procedure (RPC):
`Call call = (Call) service.createCall();`
6. Some parameters of the SOAP-client are set:

```
call.setOperationName(new QName(
    "http://www.fedora.info/definitions/1/0/api/",
    "ingest"));
call.setTargetEndpointAddress(new URL(
    "http://localhost:8180/fedora/services/management"));
call.setUsername(FEDORA_SERVER_USERNAME);
call.setPassword(FEDORA_SERVER_PASSWORD);
```

The newly created object of type `javax.xml.namespace.QName.QName` represents a *Qualified Name*, which is connected to the namespace-URI of the Fedora-API. This qualified name contains the names of the SOAP-operation (“`ingest`”). By using methods `setTargetEndpointAddress` and `setUsername` the service-endpoint of the Fedora server and the credentials for authentication are set. The call is now finished.

7. The `ingest`-method of the Fedora-API-M-SOAP-interface, besides the serialized FOXML-document, needs two more parameters – the XML-format of the object handed over and a protocol entry (*Log Message*). These are handed over during the execution of the `Call.invoke`-method:

```
String format = "foxml1.0";
String logmsg = "initial creation";
pid = (String) call.invoke(new Object[] {
    xml.toByteArray(),
    format,
    logmsg });
```

8. If the creation of the new object was successful, Fedora returns the automatically generated persistent identifier (PID) of the new object. By the entry

```
<param name="pidNamespace" value="escipub"/>
```

in Fedora's configuration file¹ the PID-prefix has been set to "escipub". This prefix is followed by a colon and a unique number. The answer to the `FedoraSOAPClient` will be e.g. "escipub:477".

After the creation of a new Fedora-object, the control flow returns to the method `startProcessInstance` of the `HomeAuthorBean`.

If the creation of the Fedora-object was successful, a new process instance of the previously chosen process definition is created:

```
ProcessDefinition processDefinition =
    graphSession.loadProcessDefinition(processDefinitionId);
ProcessInstance processInstance =
    new ProcessInstance(processDefinition);
```

Every process instance has a unique initial state. By using the call

```
TaskInstance taskInstance =
    processInstance.getTaskMgmtInstance().createStartTaskInstance();
```

the task corresponding to this initial state is created. The `AuthenticationFilter`, the `JbpmContextFilter`, and the assignment of the `ActorId` in the jBPM-context make the new task to be assigned to the right actor and the corresponding task list. The PID is saved in the process context and is therefore available to all process participants as a process variable. To make the process operations persistent, the jBPM-context is saved:

```
taskInstance.setVariable("pid", pid);
jbpmContext.save(processInstance);
```

5.3 Execution of a Fedora query

We will sketch the steps that are necessary to get the list of the author's articles by a query to Fedora:

1. The `<h:dataTable>`-tag on `home_author.jsp` expects from the `HomeAuthorBean` a list of articles of the current author. Although this list is untyped at this level, it will be possible to access all attributes of the objects in the list, as long as *Getter*-methods are available in the Java context.
2. The `HomeAuthorBean` formulates a query to the integration layer by specifying the maximum number of hits (100), the comparison operator to use `info.fedora.www.definitions._1._0.types.ComparisonOperator`², the field the query refers to ("creator"), and the value to check (the name of the current user). This query is handed over to the `FedoraSOAPClient`.

¹The Fedora configuration file can be found in `server/config/fedora.fcfig`

²This and all other Fedora types are described in [1].

The result of the query to the integration layer is an object of type `info.fedora.www.definitions._1._0.types.FieldSearchResult`.

This type encapsulates the abstract type “`resultList`”, which is of the (concrete) type `ArrayOfObjectFields`. The attributes of an `ObjectFields`-object contain Dublin Core metadata like “`creator`”, “`subject`”, and “`description`”, and Fedora object properties like the PID or the creation date (“`cDate`”) [1].


3. The method `doQuery` of the `FedoraSOAPClient` transforms the query coming from the `HomeAuthorBean` into an object of type `info.fedora.www.definitions._1._0.types.FieldSearchQuery`. A `FieldSearchQuery` consists mainly of an array of conditions; thus queries with an arbitrary number of conditions can be handled. In this case, we use only one condition. The `FieldSearchQuery` is handed over to the method `findObjects`.
4. In method `findObjects`, there is a SOAP call to the Fedora server as described above (section 5.2). But this time, there are Fedora-specific data types that are unknown to the *Axis*-library. Thus, all Fedora data types of this SOAP-call are introduced to the Axis-client as qualified name objects before the `call.invoke`-statement by the method `call.registerTypeMapping`, like for instance the data type `FieldSearchResult`¹:

```
QName qn1 = new QName(
    "http://www.fedora.info/definitions/1/0/types/",
    "FieldSearchResult");
call.registerTypeMapping(
    FieldSearchResult.class,
    qn1,
    new BeanSerializerFactory(FieldSearchResult.class, qn1),
    new BeanDeserializerFactory(FieldSearchResult.class, qn1));
```

5. The answer of the server is of type `FieldSearchresult`. It is returned to the method `doQuery`.
6. The `doQuery`-method hands on the answer to `HomeAuthorBean`.
7. Before `HomeAuthorBean` passes on the information from the integration layer to the user-interface layer the monolithic `FieldSearchResult`-object is transformed to a list of `ObjectFields`. `home_author.jsp` can access the entries of this list directly. The indexing shows that some of the Dublin Core attributes are arrays. Indeed, the Dublin Core standard has repeatable attributes.

5 The Nuts and Bolts of the Web Application

Server for publication and document storage at University of Applied Sciences Bonn-Rhein-Sieg

 eSciPub

[Logout](#)

Your are logged in as Alex

task_author.jsp

This is new article!

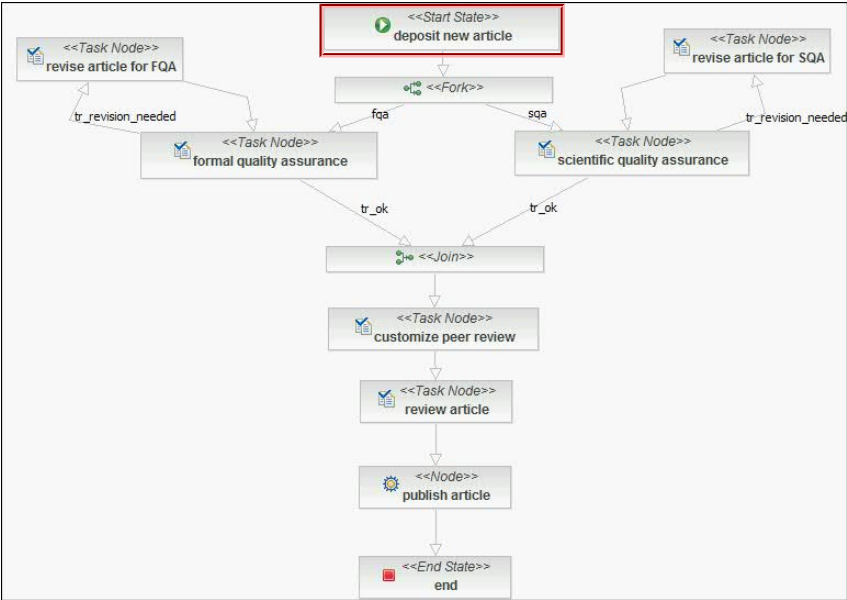
userId: 1
userName: Alex
userGroup: author
processInstanceld: 14
taskInstanceld: 31
articlePid: bislab:132

Input article metadata and upload article

PID:	bislab:132
Title:	Cold Gaseous Halos of Nearby Disk Galaxies
Author:	Alex
Keywords (English):	Disk Galaxies, Halos
Abstract:	
Responsible institution:	Department of Astronomy, University of Massachusetts,
Contributor:	Q. Daniel Wang
Language:	English
Article to upload:	<input type="text"/> <input type="button" value="Durchsuchen..."/>

Article has been uploaded!:

Uploaded file name:	Cold gaseous halos of nearby disk galaxies.pdf
Uploaded file content type:	application/pdf
Uploaded file size:	101930 Bytes



```
graph TD
    Start([<<Start State>>  
deposit new article]) --> Fork[<<Fork>>]
    Fork -- fqa --> FQA[<<Task Node>>  
formal quality assurance]
    Fork -- sqa --> SQA[<<Task Node>>  
scientific quality assurance]
    FQA -- tr_revision_needed --> FQA
    SQA -- tr_revision_needed --> SQA
    FQA -- tr_ok --> Join[<<Join>>]
    SQA -- tr_ok --> Join
    Join --> CPR[<<Task Node>>  
customize peer review]
    CPR --> RA[<<Task Node>>  
review article]
    RA --> PA[<<Node>>  
publish article]
    PA --> End([<<End State>>  
end])
```

Figure 3: Screenshot of the prototype - Input metadata and upload file

5.4 File Upload and Updating a Fedora Object

When the author sees the screen on page 13, a new empty Fedora object has already been created. The PID field has been filled automatically by the system. The author fills in the other metadata fields. Then he may want to upload a file. We will now describe the file upload functionality, because this is a feature that is not contained in jBPM.

By the statement

```
<%@ taglib uri="http://myfaces.apache.org/tomahawk" prefix="t"%>
```

in `task_author.jsp` the *Tomahawk*-tag-library is made available on this page. By using the tag

```
<t:inputFileUpload id="fileupload"
    accept="*/*"
    value="#{taskArticleBean.upFile}"
    storage="file" />
```

a file selection dialogue is displayed. The selected file of this dialogue is directly linked to the `upFile`-attribute of the `TaskArticleBean`. This attribute is of type

```
org.apache.myfaces.custom.fileupload.UploadedFile
```

from the *Tomahawk*-library (`tomahawk.jar`). Note that the file selection dialogue must be inside an HTML-form with attribute `enctype="multipart/form-data"` because the `ExtensionsFilter` responsible for the upload in the JSF-context will not recognize it otherwise. The *upload*-button is linked to the `upload`-method of the `TaskArticleBean`:

```
<h:commandButton action="#{taskArticleBean.upload}" />
```

By pressing the upload button on `task_author.jsp`, the upload method of the `TaskArticleBean` is executed and a new *HTTP-Request* is triggered. Since the *Scope* of the `TaskAuthorBean` is “Request”, this bean would be created and initialized as new by the JSF-servlet. This means that the bean’s attributes would be lost. In order not to lose this information, all attributes of the `TaskAuthorBean` that must survive the end of a request are listed in `task_author.jsp` via *Tomahawk*-tags `<t:saveState...>`.

A remarkable feature of Fedora is the fact that files cannot be imported from the local file system but only via a web-URL. That is why the file chosen for upload is first copied to the root directory of Fedora’s *Tomcat*-container. From there it can be imported by the `upload`-method of the `TaskArticleBean` using a local URL. The `TaskArticleBean` also provides important file information discovered while uploading, as for instance the *content type* and the file size. After uploading the article, the page `task_author.jsp` is rebuilt, and the information provided by the `TaskArticleBean` about the uploaded file is displayed (cf. figure 3). A copy of the file now resides in the local *Tomcat*-root-directory, e. g. in `U:/Master/fedora-2.1.1/server/jakarta-tomcat-5.0.28/webapps/ROOT/`

When the author closes the task by pressing the button “Save task and quit”, the method `saveAndCloseAuthor` of the `TaskArticleBean` is called. This method saves the metadata of

¹The classes `org.apache.axis.encoding.ser.BeanSerializerFactory` and `org.apache.axis.encoding.ser.BeanDeserializerFactory` provide objects for the serialization and deserialization of new data types.

the form on `task_author.jsp` in jBPM-process variables, so that other roles involved in the same process, e.g. the quality assurance, need not get these metadata from Fedora, but can access these process variables directly.

After that, the `TaskArticleBean` saves the metadata in the corresponding Fedora object. The PID for accessing the correct Fedora object can be read from the process variable and be handed over to the `FedoraSOAPClient`:

```
boolean success = fedoraSOAPClient.changeDC(
    articlePid, userBean.getUserName(),
    articleTitle, articleCreator, articleSubject,
    articleDescription, articlePublisher,
    articleContributor, articleDate, articleType,
    articleLanguage, articleCoverage, articleRights);
```

The method `changeDC` of the `FedoraSOAPClient` can change the metadata. Here, the new Dublin Core-data stream is built as a DOM-document: at first a new DOM-document is created with the necessary Dublin Core-namespace-attributes. Then the DC-metadata are inserted as additional nodes according to the DC-namespace-specification¹. Since Fedora creates a DC-data stream for each new object automatically, the `FedoraSOAPClient` uses the API-M-method `modifyDatastreamByValue` to save the metadata in Fedora:

```
call.setOperationName(new QName(
    "http://www.fedora.info/definitions/1/0/api/",
    "modifyDatastreamByValue"));
```

The new DOM-document containing the Dublin Core metadata is transformed to a `ByteArray` and handed over to Fedora:

```
call.invoke(new Object[] { pid, // Die PID
    "DC", // name of the data stream
    null, // altIds (no alternative identifier)
    null, // dsLabel (no change, still "ESCIPUB")
    new Boolean(true), // versioning on
    "text/xml", // MIME Type: Text
    null, // formatURI (none)
    xml.toByteArray(), // dsContent (the serialized Dublin Core)
    "A", // dsState (state - active)
    "update", // logMessage (log-entry)
    new Boolean(true) }); // force
```

Using the method `dsExists`, the `FedoraSOAPClient` has the `TaskArticleBean` find out, if the data stream with the label “ARTICLE” exists. This check is necessary because the `TaskArticleBean` is also used for reworking an existing article. Prior to saving the article in Fedora, the MIME type of the uploaded file in the local *Tomcat*-root-directory is detected:

¹s. <http://purl.org/dc/elements/1.1/>

1. By the statements

```
MultiThreadedHttpConnectionManager cManager =
    new MultiThreadedHttpConnectionManager();
HttpClient httpClient = new HttpClient(cManager);
```

a new HTTP-client of type
`org.apache.commons.httpclient.HttpClient` is created.

2. The URL of the uploaded file is contained in the string variable `localURL`. The call

```
org.apache.commons.httpclient.methods.HeadMethod head =
    new HeadMethod(localURL);
```

triggers the MIME-type detection of this file.

If the ARTICLE-Datastream exists already in the corresponding Fedora object, the Fedora-SOAPClient uses the API-M-method `modifyDatastreamByReference` - otherwise `addDatastream`. The call of the Fedora-API-M-method `modifyDatastreamByReference` is executed by the following command:

```
call.invoke(new Object[] { pid, // the PID
    "ARTICLE", // name of the data stream
    null, // altIds (no alternative identifier)
    null, // dsLabel (no change - still "ESCIPUB")
    new Boolean(true), // versioning on
    mimeType, // the MIME type identified
    creator, // formatURI (author)
    localURL, // dsLocation (local URL)
    "A", // dsState (state - active)
    "update", // logMessage (log entry)
    new Boolean(true) }); // force
```

Note that our prototype saves the name of the author in the object attribute `FormatURI`, from where this information can be retrieved more easily than by analyzing the Dublin Core-data stream. The article- and comment-lists, that are displayed to authors while reworking their articles and to quality assurers while reviewing articles, make use of this information.

After deleting the article file in the Tomcat-root-directory, the `TaskArticleBean` determines the next transition in method `close`. Then the task is finished and the user-interface layer returns to the working environment of the author.

At the lower end of the screen (cf. figure 3) the process graph is displayed; the current task is surrounded by a red rectangle. The process graph is created by the jBPM class `ProcessImageTag.java`. This class creates an HTML-table and adds the `.jpg-image`¹ of

¹This image was created by the *Graphical Process Designer* and handed over to the jBPM data base during deployment (cf. section 3.1).

the process graph as a background image of the table. By evaluating the information in file `gpd.xml`¹, the cells in the table get their red surrounding frames. The HTML-table is returned to the calling JSP. The page `task_author.jsp` is one of the JSPs of the prototype that uses this functionality. The call is made by the tag

```
<jbpm:processimage task="\${taskArticleBean.taskInstanceId}" />
```

5.5 Deployment of Process Definitions

The deployment of process definitions using the Eclipse-plugin is only possible if *eSciPub* is activated on the server. *eSciPub* contains an upload-servlet, acting as an interface between the deployment-functionality of the Eclipse-GPD-Plugin and the *jBPM*-server.

If the deployment-function is called from the environment of the process administrator, the deployment does not use this servlet, but executes the corresponding *jBPM*-function directly. The implementation of this functionality will be described now.

As a first step, the process administrator chooses a process archive file using a file selection dialogue of his workspace. When he presses the “Activate” button, an upload is performed using the Tomahawk-JSF-extension in the method `deploy` of the `AdminBean`. The selected process archive file is transformed into a byte stream by

```
ByteArrayInputStream bais =  
    new ByteArrayInputStream(_upFile.getBytes());
```

Since the process archives from GPD are compressed zip-files, the byte stream is decompressed by an instance of `java.util.zip.ZipInputStream`:

```
ZipInputStream zipInputStream = new ZipInputStream(bais);
```

Using method `parseParZipInputStream` of class

`org.jbpm.graph.def.ProcessDefinition`, the process definition of the archive can be assigned to a new `ProcessDefinition`-instance and then be deployed by method `deployProcessDefinition`:

```
ProcessDefinition processDefinition =  
    ProcessDefinition.parseParZipInputStream(zipInputStream);  
jbpmContext.deployProcessDefinition(processDefinition);
```

The deployment of the process archive is now finished.

6 Discussion

Although this work has been motivated by a scientific context, the concepts are general enough to be used by any organization that needs to manage content for internal or external purposes.

We have provided a proof of concept for the integration of an open-source digital repository into a state-of-the-art enterprise architecture.

¹This file was created by the GPD and contains information about position and size of the elements of the process graph.

The existence of a graphical user interface for the configuration of workflows may give the impression that it is easy for administrators to configure those workflows. The good news is, that adding, removing and modifying the order of tasks and the deployment of the resulting new workflows is a matter of a few minutes. The graphical representations of the workflows, that show the status of a workflow instance to the end user, are provided by the system. The bad news is, that the resulting workflow is not necessarily sound (i. e. functioning) and that no error messages are displayed during creation and deployment. At present, the only way to find out whether the new workflow is operational is testing. The documentation of a set of tasks should contain pre- and post-conditions that specify which tasks can be predecessors or successors of others. I. e. one should find out whether a static analysis can be added to jBPM that checks the soundness of workflows.

The jBPM Starters Kit¹ is a supplement to jBPM offered by JBoss. It contains a small web application, which provides tasks with very simple, string-based input fields. Based on this web application, our prototype implements custom-made tasks containing file handling and communication to Fedora. In the current prototype, there is no straightforward way of combining custom-made tasks with standard tasks. One direction to which future work should move, is the creation of a set of standard custom-made tasks that can be combined with jBPM standard tasks. Another more complex direction might be the extension of the expressivity of jBPM itself.

References

- [1] Cornell University: Fedora Development Team (ed.). Fedora types – schema fedora-types.xsd, 2005. HTML: <http://www.fedora.info/definitions/1/0/types/> (online: 08/05/2006).
- [2] Cornell University: Fedora Development Team (ed.). Fedora soap client documentation – fedora release 2.1.1, 2006. HTML: <http://www.fedora.info/download/2.1.1/userdocs/client/servlet/soapclient/index.html> (online: 07/24/2006).
- [3] JBoss Inc. (ed.). Jboss jbpmp – user guide v3.1, 2004. PDF: <http://docs.jboss.com/jbpmp/v3/userguide/> (online: 03/22/2006).
- [4] JBoss Inc. (ed.). JBoss jBPM, 2006. HTML: <http://www.jboss.com/products/jbpmp> (online: 08/30/2006).
- [5] JBoss Inc. (ed.). Jboss jbpmp starters kit v3.1.1, 2006. HTML: <http://prdownloads.sourceforge.net/jbpmp/jbpmp-starters-kit-3.1.1.zip?download> (online: 03/01/2006).
- [6] S. Kirn and R. Unland. Workflow Management mit kooperativen Softwaresystemen: State of the Art und Problemabriß, 1994. Postscript: <http://www-wi.uni-muenster.de/inst/arbber/ab29.ps> (online: 06/15/2006).

¹s. <http://www.jboss.com/products/jbpmp/downloads>

References

- [7] Max Planck Gesellschaft and Fachinformationszentrum Karlsruhe (ed.). eSciDoc – Neue Formen wissenschaftlichen Arbeitens, 2005. PDF: http://www.escidoc-project.de/fileadmin/pdf/prs_Beratssitzung_130505_final_public.pdf (online: 02/28/2006).
- [8] Max Planck Gesellschaft and Fachinformationszentrum Karlsruhe (ed.). eSciDoc – Systemsentscheidung - Entscheidung für Fedora, 2006. PDF: http://www.escidoc-project.de/fileadmin/pdf/Beiratsitzung_13012006%20Unterlagen_TOP4_public.pdf (online: 07/02/2006).
- [9] I. M. Wolfgang Dotal, Mario Jeckle and B. Zengler. *Service-orientierte Architekturen mit Web Services*. Spektrum Akademischer Verlag, München, 2005.